基于 3D 卷积神经网络的视频哈希算法 *

刘玉莹 [1]，刘宏哲 [1+]，袁家政 [2]，李　兵 [3]

(1. 北京联合大学 北京市信息服务工程重点实验室，北京 100101; 2. 北京开放大学，北京 100081; 3. 中国科学院自动化研究所模式识别国家重点实验室，北京 100190)

摘　要：随着视频分享应用和平台的蓬勃发展，视频数据正处于指数上升阶段，针对目前相似性视频检索方法中快速性和准确性仍无法满足用户要求等问题，提出了一种基于 3D 卷积神经网络的视频快速检索方法。该算法将 3D 卷积神经网络与哈希学习方法结合应用于视频数据，既能快速学习视频时空特征表示，又能极大地缩短视频检索时间。在常用视频数据集上的实验结果表明，利用所提出的方法对视频进行相似性检索性能优于当前主流方法。

关键词：深度学习；哈希算法；视频检索

中图分类号：TP　　doi: 10.19734/j.issn.1001-3695.2018.07.0664

Video hash algorithm based on 3D convolutional neural network

Liu Yuying[1], Liu Hongzhe[1+], Yuan Jiazheng[2], Li Bing[3]

(1. *Beijing Key Laboratory of Information Service Engineering Beijing*, *Beijing Union University*, *Beijing* 100101, *China*; 2. *Beijing Open University*, *Beijing* 100081, *China*; 3. *State Key Laboratory of Pattern Recognition*, *Institute of Automation*, *Chinese Academy of Science Beijing* 100190, *China*)

Abstract: With the vigorous development of video sharing applications and platforms, video data is in an exponentially rising phase. In order to solve this thorny problem that the speed and accuracy of the current similarity video retrieval methods still cannot meet the requirements of users, this paper proposes a new similarity video quick retrieval method, which combines the three-dimensional convolutional neural network with the hash learning method and apply to video data. It not only can quickly learn the video spatiotemporal feature representation but also can greatly shorten the video retrieval time. The experimental results on the set show that the similarity retrieval performance of the video using the proposed method is superior to the current mainstream methods.

Key words: deep learning; hashing method; video retrieval

0　引言

近年来，随着移动互联网技术的快速发展，图像、视频等多媒体数据呈爆炸式增长。对于互联网用户来说，从海量的视频数据中快速的检索到对自己有用或者喜欢的视频非常重要；对互联网平台来说，为用户进行个性化的视频推荐或者相关视频广告的投放能够有效地提高用户体验以及产品的销量；对视频原创者来说，能够充分对其视频进行版权保护。因此，对相似视频检索技术的研究成为基于内容的视频检索、视频个性化推荐和原创视频版权保护等应用的关键点。相似视频检索的基本思想是将查询视频与视频数据库中的视频进行近似最近邻搜索，返回与之内容相似的视频。传统的方法是先提取视频特征，然后计算查询视频特征与视频数据库中视频特征的欧氏距离，并根据距离从小到大的顺序来返回相似的视频。但是随着互联网上的视频数据的井喷式增长，传统的线性搜索方法需要的存储空间消耗大、计算复杂度高、检索速度慢。为了解决传统方法对存储空间和检索时间的上的局限，近来近似最近邻搜索技术发展迅猛，其中哈希技术作为一种代表性方法受到了广泛关注，这种方法能够将视频数据的高维特征映射到低维空间，产生简洁的二值码表示，通过计算查询视频与数据库视频二值码之间的汉明距离来进行相似度检索，能够显著降低计算开销并且提升检索性能。近几年，在许多关于计算机视觉的课题研究上运用深度学习技术都显著提升了性能，如目标检测、分类、分割等，这些任务性能的提升都归功于深度卷积神经网络在特征表示学习上的强大，但这些任务主要是针对图像作为输入进行的，因此 2D 卷积[9]能很好地对图像进行特征表示学习。而在视频作为输入的视觉任务中，2D 卷积不能及时捕获视频数据的时序信息，因此针对视频时序信息的表示，研究者们提出了 3D 卷积[10]来同时提取视频数据的时间特征和空间特征，从而保证视频特征表示的连续性。本文基于深度学习和哈希技术，提出了一种新的相似性视频快速检索方法，运用 3D 卷积神经网络[10]同时对视频进行时间特征和空间特征的提取与融合，利用哈希方法对融合后的视频时空特征进行量化编码，得到视频的哈希二值码，计算查询视频与大规模视频数据集的汉明距离，实现快速有效的视频检索。

1　相关研究

1.1　哈希学习

为了实现高效的近似最近邻搜索，哈希方法旨在将数据

收稿日期：2018-07-21；**修回日期：**2018-10-08　**基金项目：**国家自然科学基金资助项目（61871039,61571045）；国家科技支撑计划资助项目（2015BAH55F03）；北京市属高校高水平教师队伍建设创新团队建设提升计划资助项目（IDHT20170511）

作者简介：刘玉莹（1993-），女，硕士，主要研究方向为数字图像处理；刘宏哲（1971-），女（通信作者），教授，硕导，主要研究方向为语义计算、数字博物馆、分布式系统集成（liuhongzhe@buu.edu.cn）；袁家政（1970-），男，教授，博导，主要研究方向为图形图像处理、文物遗迹的数字化处理、数字博物馆等；李兵（1983-），男，副研究员，硕导，主要研究方向为视觉认知计算、多媒体内容安全.

的高维特征编码为紧凑的二值码，同时能保持原始数据之间的相似性。在数据层面可分为数据独立型[12~15]和数据依赖型[16~33]哈希方法。数据独立型哈希方法主要采用随机投影方法将数据映射为二值码，但是这种不依赖数据的哈希方法需要较多位数的二值码才能得到比较高的精度。为了解决数据独立型哈希的码长问题，数据依赖型哈希方法利用数据的属性或者数据所具有的标签信息来监督训练来生成紧凑的二值码。现有的数据依赖型哈希算法又可分为有监督哈希方法[2~5,7,8,21~31]和无监督哈希方法[1,6,16~20,32]。在无监督方法中哈希学习过程是在没有标签信息的情况下完成的。利用离散优化技术对数据的二值码进行学习，保持原始高维特征之间的相似性关系，比较典型的算法有 Yair 等人[17]提出的谱哈希(Spectral Hashing, SH)、Gong 等人[18]所提的迭代量化哈希方法(iterative quantization, ITQ)以及等人 Irie[20]提出的局部线性哈希方法(locality linear hashing, LLH)。虽然无监督哈希方法相比数据独立哈希方法的检索精度高了不少，但是缺少数据标签信息，总体的检索精度还是很难提升，因此有监督哈希方法的出现一定程度上解决了这一局限性，利用带有标签信息的数据进行哈希学习，该类哈希学习模型的目标是最小化二值码之间的距离并对增大数据之间的相似度差异，即使得相似的数据尽量靠近，不相似的数据尽量远离。在过去的几年里，由于深度卷积神经网络在各种视觉任务展现出其优越的性能，研究者们将深度学习与哈希方法结合提出了许多深度哈希方法。通过训练端到端的(convolutional neural network, CNN)[9]模型，现有的深度哈希方法能够同时学习图像表示以及在监督或者非监督的方法得到二值码。虽然现有的深度哈希方法已经取得了显著的性能，但它们大多是为图像检索设计的。相比之下，专门为视频检索设计的深度哈希方法[4~8]相对较少，因为学习视频特征表示比图像特征表示更具挑战性。由于视频提供的信息比图像提供的更加多样和复杂，在视频哈希算法设计中，研究者们主要集中在学习视频的空间特征。例如，文献[4]利用视觉注意力模型提取视频特征视觉特征，并通过深信念网络(deep belief network, DBN)融合视觉外观和视觉特征生成视频哈希，获得具有代表性的视频特征；为了融合视频的空间特征和时间特征，文献[5]提出了二值长短时记忆(binary long short-term memory, BLSTM)网络，利用 LSTM 捕获视频中的时间信息，采用二值化的 LSTM 单元在每个时间步长上输出二进制码，在处理传统的序列数据(如文本和语音)时，LSTM 具有优异的性能。然而，由于视频内容的高度多样性和复杂性，该方法在处理视频数据的泛化能力不高。现有的视频哈希方法主要集中在学习视频的感知特征，再将学习到的特征应用于图像哈希方法中去得到二值码表示，这两部分不能相互反馈，即产生的哈希码的质量在很大程度上取决于所获得的特征的质量，而哈希码没有被用来指导特征的学习。由于 3D 卷积神经网络在学习视频时空特征上的优越性，本文将视频时空特征学习与能够融入深度卷积神经网络的哈希学习相结合，提出了一种端到端的视频哈希算法，并能够快速有效地应用于相似视频检索中。

1.2　3D Convolutional Neural Networks

利用二维卷积网络对视频进行特征提取一般是对视频的每一帧图像分别利用 CNN 来进行特征学习，这种方式不会考虑时间维度的帧间运动信息，而使用 3D 卷积神经网络能更好的捕获视频中的时间和空间的特征信息。图1是 3D CNN 对图像序列（视频）采用 3D 卷积核进行卷积操作。

图 1 中进行卷积操作的时间维度为 3，即对连续的三帧图像进行卷积操作，3D 卷积是通过堆叠多个连续的帧组成一个立方体，然后在立方体中运用 3D 卷积核。在这个结构中，卷积层中每一个 feature map 都会与上一层中多个邻近的连续帧相连，以此来捕捉运动信息。例如图 1 左边，一个卷积 map 的某一位置的值是通过卷积上一层的三个连续的帧的同一个位置的局部感受野得到的。3D 卷积核只能从 cube 中提取一种类型的特征，因为在整个 cube 中卷积核的权值都是一样的，也就是共享权值，都是同一个卷积核（图中同一个颜色的连接线表示相同的权值）。因此本文可以采用多种卷积核，以提取多种特征。

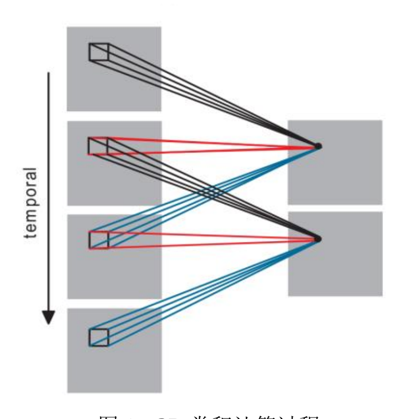


图 1　3D 卷积计算过程

Fig. 1　3D convolution alculation process

2　基于 3D CNN 的视频哈希算法

在本章中将具体介绍所提出的基于 3D 卷积神经网络的视频哈希方法模型及学习算法。

2.1　符号定义

本文使用 m 表示向量，M 表示矩阵，M^T 表示矩阵的转置。$\|\cdot\|_2$ 用来表示向量的欧氏范数，符号函数 $\mathrm{sgn}(\cdot)$ 表示如果元素为正，则返回 1，否则返回-1。假设数据集中有 n 个数据点（视频片段）$V=\{v_i\}^n_{i=1}$，其中 $v_i, v\in\mathbb{R}^D$ 表示第 i 个视频数据的 D-维特征向量；除此之外，本文还使用使用相似度矩阵 S_{ij} 表示视频数据之间的相似性，其中 $S_{ij}=1$ 表示 v_i 与 v_j 相似，$S_{ij}=0$ 表示 v_i 与 v_j 不相似。在视频检索系统中，相似度矩阵 S_{ij} 通常是用视频语义标签来构造的。本文的目标是利用深度哈希方法将具有语义标签信息的视频数据映射成哈希二值码表示，即 $b_i\in\{-1,1\}^c$，其中 c 表示二值码的长度，并且保证学到的哈希二值码能够保留视频之间的相似性 $B=\{b_i\}^n_{i=1}$，即相似视频所对应二值码之间的汉明距离尽量小，不相似视频所对应二值码之间的汉明距离尽量大，且相似视频之间的汉明距离比不相似视频之间的汉明距离更小。本文使用 $h(v_i)$ 来表示要学习的哈希函数，其中 $b_i=h(v_i)=[h_1(v_i),h_2(v_i),\ldots,h_c(v_i)]^T$。

2.2　方法模型

本方法的模型框架如图 2 所示。它是一个端到端的学习框架，主要由视频特征学习部分和哈希二值码学习部分组成，在训练过程中，各部分会相互反馈。

2.2.1 视频特征学习部分

本文的模型在文献[11]所提出的 3D CNN 模型的基础上做了一些改进，在模型的第一层卷积和第三层卷积之后增加了归一化操作，以及最后一层添加了一个哈希码学习层。请注意，图 2 中有两个 3D CNN 网络结构，这两个 3D CNN 网络具有相同的结构和相同的权重。也就是说，输入和损失函数都是基于成对对视频数据的。模型每一层的详细配置如表 1 所示。

图 2　深度视频哈希算法框架

Fig. 2　Deep video hash algorithm framework

表 1　深度视频哈希方法框架参数设置

Table 1　Parameter settings for deep video hash algorithm

Layer	Configuration
Conv1	f. 64*3*3*3;st. 1*1*1;pad. SAME;BN
Pool1	ksize:1*2*2;st. 1*2*2;pad. SAME
Conv2	f. 128*3*3*3;st. 1*1*1;pad. SAME
Pool2	ksize:2*2*2;st. 2*2*2;pad. SAME
Conv3_1	f. 256*3*3*3;st. 1*1*1;pad. SAME;BN
Conv3_2	f. 256*3*3*3;st. 1*1*1;pad. SAME
Pool3	ksize:2*2*2;st. 2*2*2;pad. SAME
Conv4_1	f. 512*3*3*3;st. 1*1*1;pad. SAME
Conv4_2	f. 512*3*3*3;st. 1*1*1;pad. SAME
Pool4	ksize:2*2*2;st. 2*2*2;pad. SAME
Conv5_1	f. 512*3*3*3;st. 1*1*1;pad. SAME
Conv5_2	f. 512*3*3*3;st. 1*1*1;pad. SAME
Pool5	ksize:2*2*2;st. 2*2*2;pad. SAME
Full6	4096
Full7	4096
Full8	Video hash code length *c*

更具体地，它包含 8 个卷积层（其中包含 5 次池化操作）和 3 个全连接层。每个卷积层描述在几个方面：f.表示卷积核的数量及其大小；st.表示卷积步长；pad.表示要添加到输入的像素数；BN 表示是否应用批量归一化操作。其中卷积核的大小均为 3*3*3，步长为 1*1*1。池化操作核的设置：除了第一层大小和步长均为 1*2*2，之后的大小和步长均为 2*2*2，这样设置是为了不过早缩减时序上的长度。除哈希码学习层外，所有层的激活函数均为 ReLU[9]，其收敛速度快且可以避免出现梯度消失问题，最后一层本文选择恒等函数作为激活函数。

2.2.2 视频哈希码学习部分

本文使用 $f(v_{i,j};\theta)\in\mathbb{R}^{c\times1}$ 表示学习到的视频特征，对应模型的输出，其中 θ 为网络模型的参数。基于 3D 卷积神经网络的视频哈希方法的目标函数定义如下：

$$\min_{B^{v_i},B^{v_j},\theta} J=-\sum_{s_{ij}\in S}(S_{ij}\Theta_{ij}-\log(1+e^{\Theta_{ij}}))$$
$$+\alpha(\|B^{v_i}-F_i\|_F^2+\|B^{v_j}-F_j\|_F^2) \qquad (1)$$
$$s.t.\quad B^{v_i},B^{v_j}\in\{-1,+1\}^c$$

其中：$F_{i,j}=f(v_{i,j};\theta),\Theta_{i,j}=0.5*F_i^T*F_j$；$B^{v_i},B^{v_j}$ 分别表示第 i,j 个视频所对应的哈希码；$B^{v_i}=sign(F_i)$；$B^{v_j}=sign(F_j)$；α 为超参数。

目标函数的第一部分 $-\sum_{i,j}^{n}(S_{ij}\Theta_{ij}-\log(1+e^{\Theta_{ij}}))$ 是具有如下定义的相似性的负对数似然函数：

$$p(S_{ij}\mid F_i,F_j)=\begin{cases}\sigma(\Theta_{ij}) & S_{ij}=1\\ 1-\sigma(\Theta_{ij}) & S_{ij}=0\end{cases}$$

其中：$\sigma(\Theta_{ij})=\frac{1}{1+e^{-\Theta_{ij}}}$，这一部分是用来保证视频特征与哈希码之间的相似性，即相似视频的哈希码尽量相似，不相似视频的哈希码尽量不同。

优化目标函数的第二部分 $\alpha(\|B^{v_i}-F_i\|_F^2+\|B^{v_j}-F_j\|_F^2)$，因为 F_i 和 F_j 是视频离散哈希码的连续替代项，保持 F_i 和 F_j 的相似性即可以保持输入视频对的相似性。

2.3　模型学习

本算法采用交替学习策略来学习 B^{v_i},B^{v_j},θ。每次先学习一个参数，其他参数固定。算法 1 简要介绍了本方法模型的整个交替学习算法，并在本节的以下内容中详细介绍该算法的推导过程。

1）学习 θ，B^{v_i},B^{v_j} 固定

当 B^{v_i},B^{v_j} 固定时，本文使用反向传播算法来学习网络模型的参数 θ。根据现有的深度学习方法[9],本文利用随机梯度下降法反向传播梯度来学习 θ。更具体地，在每次迭代中，本文从训练集采样一小部分的视频数据，然后基于采样数据执行本文的学习算法。特别地，对于每个采样点组，本文首先计算以下梯度：

$$\frac{\partial J}{\partial F_i}=0.5*\sum_{j=1}^{n}(\sigma(\Theta_{ij})F_j-S_{ij}F_j)+2\alpha(F_i-B^{v_i})$$
$$\frac{\partial J}{\partial F_j}=0.5*\sum_{i=1}^{n}(\sigma(\Theta_{ij})F_i-S_{ij}F_i)+2\alpha(F_j-B^{v_j}) \qquad (2)$$

利用所求的 $\frac{\partial J}{\partial F_i},\frac{\partial J}{\partial F_j}$ 通过复合求导链式法则再求 $\frac{\partial J}{\partial\theta}$，然后利用反向传播算法即可更新参数 θ。

2）学习 B^{v_i},B^{v_j}，θ 固定

当参数 θ 固定时，式（1）将重新定义如下：

$$\max_{B^{v_i},B^{v_j}} tr[\alpha((B^{v_i})^T F_i+(B^{v_j})^T F_j)] \qquad (3)$$

则哈希码的更新依赖于其连续替代项：

$$B^{v_i}=sign(\alpha(B^{v_i})^T F_i),B^{v_j}=sign(\alpha(B^{v_j})^T F_j) \qquad (4)$$

算法 1　基于 3D 卷积神经网络的视频哈希算法

输入：视频数据集 V 以及相似度矩阵 S。

输出：模型的网络参数 θ，以及输入的视频所对应的哈希二值码 B。

初始化：初始化模型网络参数 θ，mini-batch size=10，迭代 iteration =5000。

```
Repeat
for iteration=1,2,3,..,5000 do
  从 V 中随机取样视频数据来构造一个 mini-batch
  for 每个在 mini-batch 的采样视频组( v_i,v_j )，
    根据前向传播算法计算 F_{i,j} = f(v_{i,j};θ)，并根据式（2）计算对应的梯度并根据反向传播算法更新模型网络参数 θ。
  end for
  根据式（4）得到输入视频所对应哈希二值码
Until iteration =5000
```

3　实验

在这一部分中，本文在常用的视频数据集上与其他最先进的方法进行了比较，验证了所提方法的有效性。本实验是在 Nvidia Titan X GPU 服务器上使用开源的深度学习框架 Tensorflow 实现的。

3.1　数据集和评价指标

UCF101 数据集包括 101 个动作类别，13 320 个实际动作视频片段，UCF101 中大多数视频的剪辑持续时间小于 10 s，在实际实验中本文选择了其中的 9 537 个视频片段作为训练集，剩余 3 783 个片段作为测试集。

对于基于哈希的检索方法，汉明距离排序和哈希查找是两种广泛使用的检索性能评估方法。在本实验对比中也采用这两种性能来评估本文方法和其他的 baselines。汉明距离排序[34]检索评估方法是将数据库中的视频与给定查询视频的汉明距离按从小到大的顺序进行排列，平均均值精度(mean average precision, mAP)[33]是衡量汉明距离排序准确性的常用指标。哈希查找是返回数据库中离查询视频在某个汉明半径以内的所有视频，而精确—召回率曲线是用来衡量哈希查找方法准确性的通用指标。

3.2　实验参数设置

根据数据集中视频语义级标签来构建相似度矩阵。在训练阶段，网络的输入是由两个输入视频的帧集组成的帧内对；而在检索阶段，输入是单个视频的帧集。每个帧集包含从视频中随机选择的帧数 k，在实验中，本文将 k 设为 16，帧集中的每个帧被调整为 112×112。为了评估不同长度的哈希码的性能，本文将二进制码的长度分别设为 16、32 和 64。目标参数中 α 值设置为 1。

3.3　Baseline

本文将所提出的方法与目前最先进的视频检索 baselines 进行了比较，包括三种传统哈希方法：(from image hashing to video hashing, FIHTV)[1]、(video hashing via structure learning, VHSL)[3]和(submodular video hashing, SVH)[2]，以及三种深度哈希方法： (deep video hashing, DVH)[7]、(video hashing based on appearance and attention features fusion via DBN ,DBNVH)[4]和(unsupervised deep video hashing with balanced rotation,BRVH)[6]。在视频检索性能对比实验中，尽量按照对比论文中作者提出的参数设置复现方法。对于与传统视频哈希方法进行比较侧重于哈希学习方法上，而针对深度视频哈希学习的算法，性能比较的侧重点在视频特征学习部分。本文还将提出的方法与图像深度哈希方法进行对比：(deep supervised hashing for fast image retrieval, DSH) [31]，(deep learning of binary hash codes for fast image retrieval, DLBHC)[30]等。这一部分实验中，将图像哈希码学习部分的方法实验代码合理改编并能够应用于视频的哈希码学习，主要用于对比所提出哈希码学习方法。

3.4　实验结果与分析

3.4.1 汉明排序

表 2　本文方法及其他方法在汉明排序上的最佳 mAP

Table 2　Best map of this method and other methods in Hamming ranking

任务	方法	UCF101 16 bits	UCF101 32 bits	UCF101 64bits
V—>V	FIHTV	0.3651	0.3725	0.4213
	VHSL	0.3985	0.4202	0.4758
	SVH	0.4564	0.4787	0.4886
	UVH	0.5041	0.5472	0.5830
	DBNVH	0.5612	0.6253	0.6441
	BRVH	0.6525	0.6826	0.7189
	DSH*	0.7274	0.7458	0.7963
	DLBHC*	0.7141	0.7199	0.7418
	本方法	**0.7832**	**0.7914**	**0.8256**

在表 2 中展示了本文方法及其他方法在汉明排序上的 mAP，“V—>V 表示查询视频对数据库视频的检索。从表 2 中可知，相比于传统视频哈希方法，本文所采用的哈希码学习方法更为高效，检索性能提升明显。为了进一步验证基于 3D CNN 的视频哈希算法的有效性，本文利用在 Sport-1M 数据集上预先训练的深层网络来提取视频时空特征，与其他深度视频哈希算法学习视频特征的方式相比，3D 卷积神经网络学习到视频特征更能有效地表示视频内容，融入到哈希学习的部分进行相互反馈能保持视频特征的相似性，因此能得到更高的检索精度。

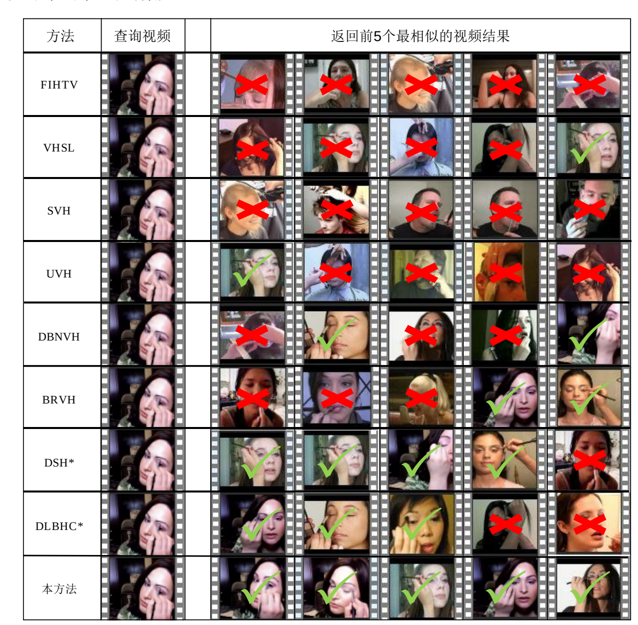


图 3　查询结果(32 bit)

Fig. 3　Query results (32 bit)

为了验证所提方法中目标函数的有效性，本文将所求得的深度视频哈希模型与其他视频哈希方法进行性能比较，如图 3 所示，将查询视频输入到模型中产生 32 bit 的哈希二值码，并利用汉明距离对数据库中可能的相似视频进行排序。由于空间限制，本文只返回结果中前五个最相似的视频结果，其中，绿色对勾表示返回的结果与查询视频相似，反之红色叉号表示返回的结果与查询视频并不相似。可以观察到，本文所提出的方法得到了最好的效果，从而验证了所提的深度

视频哈希学习方法的高效性。

3.4.2 哈希查找

在哈希查找协议中，本文可以计算出给定任何汉明半径的返回点的精确率和召回率，通过将汉明半径从 0 变到 d，步长为 1，就可以得到精确—召回率曲线。图 4 显示了本文方法和其他 baselines 方法在 UCF101 数据集上哈希码长度为 16 的精确—召回曲线。可以发现本文提出的方法获得最佳的性能。

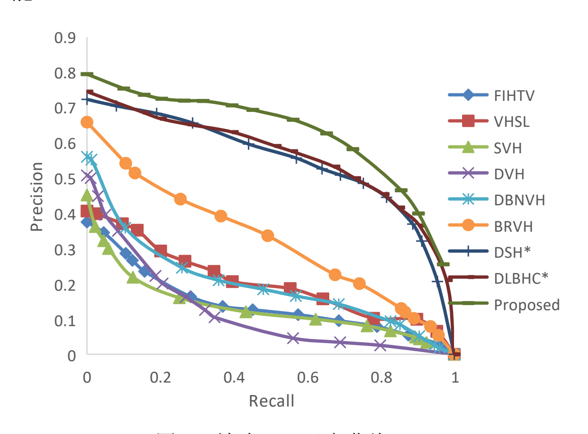


图 4　精确—召回率曲线(16 bit)

Fig. 4　Precision-recall rate curve (16 bit)

4　结束语

本文通过 3D 卷积神经网络学习视频的时空特征，并将学习到的特征融入到哈希算法中，根据为视频相似度保持任务所设计的目标函数进行大量的训练能够得到紧凑并保持视频时空特征相似性的二值码，摆脱了传统视频特征手工设计的局限，大幅降低了大规模视频数据集特征存储的空间，因此在视频检索应用中，该方法既能够加快对相似视频的检索速度，又能提高检索精度。

参考文献：

[1] Li Weng, Bart P. From image hashing to video hashing [C]// Proc of International Conference on Advances in Multimedia Modeling. [S.l.]:Springer-Verlag, 2010: 662-668.

[2] Cao Liangliang, Li Zhenguo, Mu Yadong, *et al.* Submodular video hashing: a unified framework towards video pooling and indexing [C]// Proc of ACM International Conference on Multimedia. 2012: 299-308.

[3] Ye Guangnan, Liu Dong, Wang Jun, *et al.* Video Hashing via Structure Learning [C]// Proc of IEEE International Conference on Computer Vision. 2014: 2272-2279.

[4] Sun Jiande, Liu Xiaocui, Wan Wenbo, *et al.* Video hashing based on appearance and attention features fusion via DBN [J]. Neurocomputing, 2016, 213: 84-94.

[5] Zhang Hanwang, Wang Meng, Hong Richang, *et al.* Play and rewind: optimizing binary representations of videos by self-supervised temporal hashing [C]// Proc of ACM on Multimedia Conference. 2016: 781-790.

[6] Wu Gengshen, Liu Li, Guo Yuchen, *et al.* Unsupervised deep video hashing with balanced rotation [C]// Proc of the 26th International Joint Conference on Artificial Intelligence. 2017: 3076-3082.

[7] Liong V E, Lu Jiwen, Tan Y P, *et al.* Deep video hashing [J]. IEEE Trans on Multimedia, 2017, 19 (6): 1209-1219.

[8] Wang Wulin, Sun Jiande, Liu Ju. A memorability based method for video hashing [C]// Proc of IEEE International Conference on Communication Technology. 2016: 309-313.

[9] Alex K, Ilya S, Hinton H G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]:Curran Associates Inc, 2012: 1097-1105.

[10] Tran D, Bourdev L, Fergus R, *et al.* Learning spatiotemporal features with 3d convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4489-4497.

[11] Ji Shuiwang, Xu Wei, Yang Ming, *et al.* 3D convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35 (1): 221-231.

[12] Li Ping, Shrivastava A, Moore J. Hashing algorithms for large-scale learning [J]. Nips, 2011: 2672-2680.

[13] Shrivastava A, Li Ping. Densifying one permutation hashing via rotation for fast near neighbor search [C]// Proc of International Conference on International Conference on Machine Learning. 2014: I-557.

[14] Shrivastava A, Li Ping. Asymmetric LSH (ALSH) for sublinear time Maximum Inner Product Search (MIPS) [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]:MIT Press, 2014: 2321-2329.

[15] Alexandr A, Piotr I, Thijs L, *et al.* Practical and optimal LSH for angular distance [J]. Computer Science, 2015.

[16] Liu Wei, Wang Jun, Kumar S, *et al.* Hashing with Graphs [C]// Proc of International Conference on, Machine Learning. 2011: 1-8.

[17] Yair W, Antonio T, Rob F. Spectral hashing [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]:Curran Associates Inc, 2008: 1753-1760.

[18] Gong Yunchao, Lazebnik S, Gordo A, *et al.* Iterative quantization: a Procrustean approach to learning binary codes for large-scale image retrieval. [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35 (12): 2916-2929.

[19] Jiang Qingyuan, Li Wujun. Scalable graph hashing with feature transformation [C]// Proc of International Conference on Artificial Intelligence. [S.l.]:AAAI Press, 2015: 2248-2254.

[20] Irie G, Li Zhengguo, Wu Xiaoming, *et al.* Locally linear hashing for extracting non-linear manifolds [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]:IEEE Computer Society, 2014: 2123-2130.

[21] Liu Wei, Wang Jun, Ji Rongrong, *et al.* Supervised hashing with kernels [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2012: 2074-2081.

[22] Karen S, Andrew Z. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014.

[23] Wang Jun, Kumar S, Chang S F. Semi-supervised hashing for scalable image retrieval [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition.2010: 3424-3431.

[24] Jiang Qingyuan, Li Wujun. Deep cross-modal hashing [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2017: 3270-3278.

[25] Liong V E, Lu Jiwen, Tan Y P, *et al.* Cross-modal deep variational hashing [C]// Proc of IEEE International Conference on Computer Vision. 2017: 4097-4105.

[26] Xia Rongkai, Pan Yan, Lai Hanjiang, *et al.* Supervised hashing for image retrieval via image representation learning [C]// Proc of AAAI International Conference on Artificial Intelligence. 2014.

[27] Lai Hanjiang, Pan Yan, Liu Ye, *et al.* Simultaneous feature learning and

hash coding with deep neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 3270-3278.

[28] Liong V E, Lu Jiwen, Wang Gang, *et al.* Deep hashing for compact binary codes learning [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2015: 2475-2483.

[29] Li Wujun, Wang Sheng, Kang Wangcheng. Feature learning based deep supervised hashing with pairwise labels [C]// Proc of International Joint Conference on Artificial Intelligence. [S.l.]:AAAI Press, 2016: 1711-1717.

[30] Lin K, Yang H F, Hsiao J H, *et al*. Deep learning of binary hash codes for fast image retrieval [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]:IEEE Computer Society, 2015: 27-35.

[31] Liu Haomiao, Wang Ruiping, Shan Shiguang, *et al*. Deep supervised hashing for fast image retrieval [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2016: 2064-2072.

[32] Lin K, Lu Jiwen, Chen Chusong, *et al*. Learning compact binary descriptors with unsupervised deep neural networks [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2016: 1183-1192.

[33] Liu Wei, Mu Cun, Kumar S, *et al.* Discrete graph hashing [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]:MIT Press, 2014: 3419-3427.

[34] Mohammad N, David J F, Salakhutdinov R. Hamming distance metric learning [C]//Advances in Neural Information Processing Systems. 2012: 1061-1069.